Why and how is this a “related document”? :
Semantics-based analysis of and navigation through
heterogeneous text corpora

Final project report
Bettina Berendt*
2008-02-18

Summary. In this project, a software was developed that (a) demonstrates the abilities of
software developed in and around PASCAL (more specifically, Textgarden), (b) complements
the available software by new features, and (c) demonstrates the real-life value of the resulting
text-analysis procedures by applying them to the search for blogs and news. The software is

Motivation and scientific background
The motivation and scientific background are described in the project proposal.

Results: Software components
The following figure shows the states and state transitions of document processing:

In accordance with the project proposal (as modified in the email exchange in response to the
acceptance and funding decisions), the final product consists of the following components:

* Project responsible. The project was carried out by a team also involving Marko Grobelnik, Dunja Mladenić
and Blaž Fortuna, all Jožef Stefan Institute, Ljubljana, Slovenia. Many thanks to Daniel Trümper!
**Back-end / server-side module** (main function in figure above: search)
- Software component for crawling / extracting the specified archives
  - A crawler/wrapper has been developed to extract content from English-language news (Yahoo! news, http://news.yahoo.com/) and English- and German-language blogs. The latter is generic; it has been tested on blogs from a number of key English and German sites (see below, “Tests”). It operates via Blogdigger (http://www.blogdigger.com).
- Software component for preprocessing the documents
  - Non-English documents are translated by a self-developed translator that employs Babelfish (http://babelfish.altavista.com/).
  - All documents are then transformed into a bag-of-words representation using Textgarden’s Txt2Bow software (in a library version provided directly to us from the developing team at JSI, Ljubljana).
  - Named Entities are recognized and extracted using GATE (http://gate.ac.uk/).
- Software component for enriching metadata by the learned features
  - For each document, identifier, name, date, bag-of-words and bag-of-named-entities are stored in a database.
  - For each document pair, the cosine similarity between the bag-of-words vector and between the bag-of-named-entities vector are stored.
- Software component for interfacing to the database to provide functionality for multidimensional analysis
  - The database described in the previous step is accessed via SQL and Java interfaces to SQL in the appropriate modules of the tool.
- Integration of all of the above into the back-end module
- Output: Database and XML archives

**Front-end / client-side module** (functions in figure above: global and local analysis)
- UI software component for specification of corpus and document choice and criteria of interest
  - via Web interface
- UI software component for handling/exchanging the data that are present at the client side
  - by download via the Web interface
- UI software component for global analysis
  - This is done by Ontogen, which is made available on the Web interface (redirected to its original download source, http://ontogen.ijs.si/). Ontogen supports document analysis prior to ontology learning via the integrated DocumentAtlas component.
  - It requires the user to have a current .NET environment installed.
- UI software component for local analysis
  - This software is available for download via the Web interface.
  - It requires the user to have a JRE 1.6 installed.
- Integration of all of the above into the front-end module

**Tests** have been carried out on the following blog sites
- http://www.huffingtonpost.com (English)
- http://www.dailykos.com (English)
- http://michellemalking.com (English)
- http://pajamasmedia.com (English)
Parameters have been tuned according to these experiments.

The server-side components are installed at http://www.cs.kuleuven.be/~berendt/PORPOISE; and the client-side components can be downloaded from there, together with a demonstration corpus of 1129 documents.

**Results: User interaction and added value for information search**

The project’s motivation and results and an overview of user interaction are described in the slideset of the presentation given at the PASCAL Symposium in Bled on 29th January, 2008. The slides are available at http://www.cs.kuleuven.be/~berendt/PORPOISE/porpoise_2008_01_29.ppt; the whole presentation will also soon be available on http://www.videolectures.net.

The following figure shows examples of global analysis (top) and local analysis (bottom).
Why and how is this a “related document”? ...